ANALISIS SENTIMEN TERHADAP BAKAL CALON PRESIDEN 2024 DENGAN ALGORITMA MULTINOMIAL NAÏVE BAYES DAN OVERSAMPLING SMOTE

Hary Permana, Yulison Herry Chrisnanto, Herdi Ashaury
Informatika, Universitas Jenderal Achmad Yani Cimahi
Jl. Terusan Jend. Sudirman, Cibeber, Kec. Cimahi Sel., Kota Cimahi, Jawa Barat 40531
hary.permana@student.unjani.ac.id

ABSTRAK

Pemilihan presiden 2024 menjadi salah satu bahasan yang ramai dibicarakan oleh masyarakat Indonesia, khususnya di media sosial Twitter. Pengguna Twitter menuliskan opini mereka melalui tweet terhadap bakal calon presiden yang akan maju pada pilpres 2024 nanti. Dari opini tersebut dapat dilakukan penelitian untuk mengetahui sentimen masyarakat Indonesia terhadap tokoh bakal calon presiden 2024 dengan melakukan analisis sentimen. Salah satu algoritma yang dapat digunakan untuk melakukan analisis sentimen adalah Multinomial Naïve Bayes, algoritma ini menunjukkan performa yang baik dalam melakukan analisis sentimen. Namun, dalam melakukan analisis sentimen biasanya terdapat masalah pada ketidakseimbangan kelas. Untuk mengatasi hal tersebut, penelitian ini menggunakan teknik oversampling SMOTE untuk menyeimbangkan kelasnya. Pengumpulan data dilakukan dengan *crawling* data tweets menggunakan *library snscrape*. Selanjutnya, tahapan *preprocessing* dilakukan untuk menghilangkan kata yang duplikat dan tidak relevan. Karena data tweets tidak mempunyai label maka akan dilakukan pelabelan menggunakan *lexicon InSet*. Setelah itu dilakukan ekstraksi fitur menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) dengan nilai 1-gram dan 2-gram. Hasilnya menunjukkan penggunaan oversampling SMOTE dapat meningkatkan performa klasifikasi ketika kelas data tidak seimbang, yang ditunjukkan dengan adanya peningkatan performa berkisar antara 6-20% pada setiap dataset.

Kata kunci : *analisis sentimen, bakal calon presiden, inset, multinomial naïve bayes, smote*

1. PENDAHULUAN

Pemilihan umum tahun 2024 membuka lembaran sejarah baru untuk bangsa Indonesia karena akan berlangsungnya pemilihan umum serentak seluruh Indonesia yang meliputi pemilihan presiden, pemilihan wakil rakyat DPR RI, DPD RI, pemilihan gubernur, pemilihan bupati/walikota, serta DPRD provisi/kota[1]. Pemilihan presiden di Indonesia diatur dalam Undang-Undang Nomor 42 Tahun 2008 yang dilaksanakan setiap 5 tahun sekali, dan apabila seorang presiden telah menjabat selama 2 periode berturut-turut tokoh tersebut tidak boleh mencalonkan kembali menjadi presiden. Dari aturan tersebut maka presiden yang telah menjabat selama 2 periode berturut-turut tidak dapat mencalonkan kembali dalam pemilihan presiden 2024. Antusiasme dari pemilihan presiden (pilpres) sudah mulai terasa dari saat ini, salah satu contohnya adalah mulai terbentuknya poros koalisi baru dan deklarasi dari ketua umum partai yang mencalonkan sebagai calon presiden 2024[2].

Beberapa lembaga survei telah melakukan survei elektabilitas terhadap beberapa tokoh publik yang memiliki peluang untuk maju di Pilpres 2024. Pada penelitian ini diambil 2 hasil survei tentang bakal calon presiden 2024, lembaga Charta Politika yang merilis hasil survei yang telah dilakukan terkait elektabilitas calon presiden 2024, dari 10 nama terdapat 3 nama yang mendapatkan nilai elektabilitas tertinggi yakni Ganjar Pranowo dengan 38,2% disusul elektabilitas Prabowo Subianto sebesar 31,1% dan Anies Baswedan sebesar 23,9%[3]. Selanjutnya hasil survei yang dilakukan oleh Saiful Munjani Research and Consulting (SMRC) mencatat elektabilitas dari Ganjar Pranowo sebesar 24,6%, Prabowo Subianto sebesar 17,2%, dan Anies Baswedan sebesar 10,2% [4]. Dari hasil survei tersebut muncul 3 nama yang berpotensi maju sebagai calon presiden dalam pilpres 2024 nanti, yakni Ganjar Pranowo, Anies Baswedan, dan Prabowo Subianto karena selalu memiliki elektabilitas yang cukup tinggi.

Seiring berkembangnya waktu penggunaan media sosial, khususnya Twitter di Indonesia terus meningkat, Pada bulan Januari 2023 Indonesia memiliki 167 juta pengguna media sosial. Untuk media sosial Twitter, Indonesia tercatat memiliki sekitar 24 juta pengguna pada awal tahun 2023[5]. Twitter adalah sebuah layanan jejaring sosial yang dapat digunakan oleh penggunanya untuk mengirim dan membaca pesan berbasis teks. Pengguna Twitter umumnya mengungkapkan pikiran, perasaan, dan pendapat mereka tentang isu terkini yang menjadi *trending topic* seperti kebijakan pemerintah, isu sosial-politik, dan bencana alam.

Analisis sentimen merupakan salah satu tipe dari penelitian data mining, yang berfokus untuk menganalisa, mengekstrak, dan mengolah data tekstual secara otomatis untuk mendapatkan informasi yang terkandung dalam sebuah kalimat opini. Salah satu metode yang dapat digunakan untuk mengklasifikan suatu opini kedalam opini positif atau negatif adalah dengan menggunakan Naïve Bayes[6]. Naïve Bayes merupakan salah satu metode probabilistik yang menganggap bahwa setiap atribut data tidak bergantung satu sama lain. Algoritma Naïve Bayes memiliki yang biasa digunakan dalam text mining adalah Multinomial Naïve Bayes (Multinomial

NB) yang menganggap independensi antara kemunculan kata dalam sebuah dokumen, tanpa memperhatikan urutan kata dan konteks informasi dalam dokumen[6].

Berdasarkan penelitian terdahulu yang berjudul "*A Comparative Analysis of Machine Learning Classifiers for Twitter Sentiment Analysis*" yang menggunakan beberapa varian algoritma seperti Multinomial Naïve Bayes, Bernoulli Naïve Bayes, dan SVM. Hasilnya menunjukkan bahwa algoritma Multinomial NB mendapatkan akurasi yang tinggi dibandingkan SVM dan Bernoulli NB[7]. Penelitian selanjutnya tentang analisis sentimen menggunakan Naïve Bayes, Bernoulli NB, dan Multinomial NB yang menggunakan 2 dataset, yakni untuk dataset pertama dilabeli berdasarkan emosional seseorang dan dataset kedua dilabeli dengan positif dan negatif. Hasilnya menunjukkan Multinomial NB lebih unggul dalam hal akurasi untuk kedua dataset tersebut dibandingkan Bernoulli NB dan Naïve Bayes. Multinomial NB mendapatkan akurasi sebesar 91.6% pada dataset pertama sedangkan dataset kedua sebesar 87.6%[8]. Hal tersebut menunjukkan algoritma Multinomial NB menunjukkan performa yang cukup baik dalam analisis sentimen.

Ketidakseimbangan kelas menjadi isu yang banyak muncul dan mempunyai pengaruh yang signifikan pada *data mining*[9]. Pada klasifikasi dengan dua kelas (biner), tingkat ketidakseimbangan kelas dapat ditunjukkan dengan rasio ukuran sampel suatu kelas lebih banyak dibanding kelas lainnya. Berdasarkan beberapa penelitian, menunjukkan bahwa faktor penyebabnya melibatkan distribusi kelas, jumlah sampel yang terbatas, dan pemisahan kelas[9]. Untuk mengatasi masalah tersebut, penelitian ini akan menggunakan metode *oversampling* SMOTE, yang dimana ini membuat data sintesis yang menyerupai data pada kelas minoritas[10]. Penelitian [11] menunjukkan penggunaan SMOTE pada data sentimen yang tidak seimbang akan meningkatkan performa model sebesar 12%.

Berdasarkan pemaparan yang telah dijelaskan, penelitian ini akan melakukan analisis sentimen terhadap bakal calon presiden 2024 menggunakan algoritma Multinomial Naïve Bayes dengan metode oversampling SMOTE. Data yang digunakan pada penelitian ini merupakan data opini dari media sosial Twitter yang di *crawling* menggunakan *library snscrape*.

2. TINJAUAN PUSTAKA

2.1. *Text Mining*

Text mining adalah suatu teknik dalam kecerdasan buatan yang sebelumnya mengubah data tidak terstruktur menjadi data yang terstruktur dengan *Natural Language Processing* agar dapat meningkatkan analisis menggunakan pembelajaran mesin[12]. Tujuannya adalah untuk mendapatkan pengetahuan yang sebelumnya tidak diketahui namun berpotensi berguna sebagai pengetahuan lebih lanjut dari sebuah data teks yang tidak terstruktur atau semi-struktur. *Text mining* bisa digunakan dalam berbagai bidang seperti analisis sentimen, ekstraksi entitas, dan klasifikasi teks.

2.2. Analisis Sentimen

Analisis sentimen merupakan salah satu jenis klasifikasi teks yang mengklasifikasi teks berdasarkan dari orientasi sentimen yang dikandungnya[13]. Analisis sentimen juga adalah tipe dari data mining yang menangani opini seseorang melalui *Natural Language Programming* (NLP), komputasi linguistik, dan analisis teks[14]. Terdapat dua pendekatan untuk mengestrak sentimen dari opini yang diberikan dan mengklasifikasi hasilnya sebagai positif atau negatif, yaitu dengan pendekatan berbasis *Lexicon* dan *Machine Learning Approach*[14]. Pendekatan berbasis *Lexicon* membutuhkan *lexicon* yang sudah ditentukan sebelumnya sementara *Machine Learning Approach* sudah secara otomatis melakukan klasifikasi opini, namun membutuhkan data latih. Analisis sentimen merupakan sebuah proses yang mendeteksi polaritas konstektual dari suatu teks[13]. Ini menentukan apakah suatu teks yang diberikan positif atau negatif.

2.3. *Multinomial Naïve Bayes*

Multinomial Naïve Bayes (Multinomial NB) merupakan varian dari algoritma *Naïve Bayes* yang cocok untuk klasifikasi teks atau dokumen[15]. *Multinomial NB* termasuk ke dalam jenis *supervised learning* untuk melakukan proses klasifikasi dokumen berdasarkan nilai probabilitas suatu kelas pada suatu dokumen[6]. Kata *multinomial* pada *multinomial* NB mengacu kepada distribusi multinomial. Distribusi multinomial adalah distribusi probabilitas yang digunakan pada eksperimen dengan dua variabel atau lebih. Perbedaan antara Naïve Bayes dan Multinomial NB adalah bahwa Naïve Bayes bekerja berdasarkan *conditional probability* (independensi kemunculan dipertimbangkan), sedangkan Multinomial NB bekerja berdasarkan distribusi multinomial[8]. Untuk menghitung probabilitas dari dokumen d yang terletak pada kategori c dapat menggunakan persamaan[16]:

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k|c) \qquad (1)$$

$P(t_k|c)$ merupakan probabilitas kondisional dari t_k pada dokumen dengan kategori c. $P(c)$ adalah probabilitas *prior* dari kategori c. Persamaan probabilitas *prior* memiliki persamaan[16]:

$$P(c) = \frac{N_c}{N} \qquad (2)$$

N_c merupakan jumlah kelas c yang terdapat pada dokumen dan N merupakan jumlah seluruh dokumen. Probabilitas kondisional untuk menghitung kata pada dokumen dengan kategori c menggunakan persamaan:

$$P(t_k|c) = \frac{w_k + 1}{|V| + N'} \qquad (3)$$

w_k merupakan kemunculan jumlah t_k pada dokumen latih dengan kategori c. Penambahan angka satu digunakan sebagai Laplacian Smoothing untuk

menghindari probabilitas nol. N' merupakan jumlah total kata atau *term* yang muncul dalam suatu dokumen latih c. $|V|$ adalah jumlah seluruh kata unik pada suatu dokumen (*vocabulary size*) pada data latih.

2.4. *Lexicon-Based Approach*

Pendekatan berbasis *Lexicon* dikenal sebagai metode *unsupervised learning*. Metode ini hanya bergantung pada kamus dan tidak memerlukan data latih[17]. *Lexicon* sentimen merupakan daftar dari kata-kata yang telah diberikan label positif atau negatif berdasarkan orientasi semantiknya. Ada beberapa contoh leksikon salah satunya seperti LIWC (*Linguistic Inquiry and Word Count*), GI (*General Inquirer*) yang mengkategorikan kata menjadi suatu positif atau negatif berdasarkan konteks orientasi semantik bebas. Salah satu keunggulan utama dari pendekatan berbasis *Lexicon* adalah independensi domainnya[14].

Salah satu alat analisis sentimen yang berbasis *lexicon* adalah *Indonesia Sentiment Lexicon* (InSet). InSet merupakan *lexicon* berbahasa Indonesia yang dapat digunakan di microblog. InSet memiliki data kamus sekitar 10.000 kata, dengan 3.609 kata dengan sentimen positif, dan 6.609 kata dengan sentimen negatif. Kata pada kamus tersebut dikumpulkan dari media sosial Twitter. Setiap kata memiliki bobot yang berkisar dari angka -5 (sangat negatif) sampai dengan 5 (sangat positif). Lexicon InSet juga telah dievaluasi untuk membandingkannya dengan *lexicon* bahasa Indonesia Vania *Lexicon* dan beberapa *lexicon* bahasa Inggris yang diterjemahkan seperti *Senti WordNet*, *Liu Lex*, dan *AFINN*. Hasilnya, akurasi *lexicon* InSet lebih unggul dibanding beberapa *lexicon* tersebut yakni sebesar 65,78%[18].

2.5. *Synthetic Minority Oversampling Technique* (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) merupakan teknik *oversampling* yang digunakan untuk menangani data yang tidak seimbang dengan cara membuat data sampel sintesis untuk kelas minoritas[10]. Ketidakseimbangan data muncul ketika jumlah sampel dalam kelas mayoritas jauh lebih banyak dibandingkan dengan jumlah sampel kelas minoritas. SMOTE menunjukkan hasil yang efektif dalam meningkatkan kinerja klasifikasi ketika data tidak seimbang[19]. Dalam proses pembuatan sampel sintesis, SMOTE memiliki perbedaan untuk dataset yang bersifat numerik dan dataset yang bersifat nominal (kategorikal). Untuk data numerik, SMOTE menggunakan *Euclidean Distance* untuk menghitung jarak antar *instance*. Sedangkan untuk data kategorikal, SMOTE menggunakan *Value Diffence Metric* (VDM) yang dimodifikasi untuk menghitung jarak antar instance[10].

2.6. *Term Frequency – Inverse Document Frequency* (TF-IDF)

TF-IDF merupakan proses transformasi data dari data teks ke dalam data numerik dengan pembobotan setiap kata atau fitur[20]. TF-IDF ini digunakan karena dapat memberikan peningkatan akurasi terhadap analisis sentimen[21]. *Term Frequency* (TF) memiliki artian frekuensi kemunculan suatu kata yang terdapat dalam suatu dokumen[22]. Suatu kata yang memiliki nilai TF tinggi memiliki arti penting dalam dokumen. Untuk menghitung TF dapat menggunakan persamaan berikut:

$$TF(t,d) = \frac{n(t,d)}{N(d)} \quad (4)$$

Document Frequency (DF) ialah frekuensi kemunculan suatu kata dalam seluruh dokumen. Suatu kata yang memiliki nilai DF tinggi tidak memiliki arti penting karena kata tersebut muncul di seluruh dokumen[22]. *Inverse Document Frequency* (IDF) adalah kebalikan dari DF yang berguna untuk mengukur seberapa penting suatu kata dalam seluruh dokumen. Nilai IDF yang tinggi menunjukkan kata tersebut hanya muncul dalam sedikit dokumen dan meningkatkan kepentingan. Untuk menghitung IDF dapat menggunakan persamaan berikut:

$$IDF_t = \log(\frac{N}{dt(t)}) \quad (5)$$

Lalu, pembobotan dengan TF-IDF dapat dihitung menggunakan persamaan:

$$TF - IDF_{t,d} = TF_{t,d} \times IDF_t \quad (6)$$

Keterangan:
$TF_{t,d}$: Frekuensi kemunculan kata t dalam suatu dokumen
IDF_t *: Inverse document frequency* pada kata t
$TF - IDF_{t,d}$: Bobot nilai d terhadap kata t

2.7. *Data Mining*

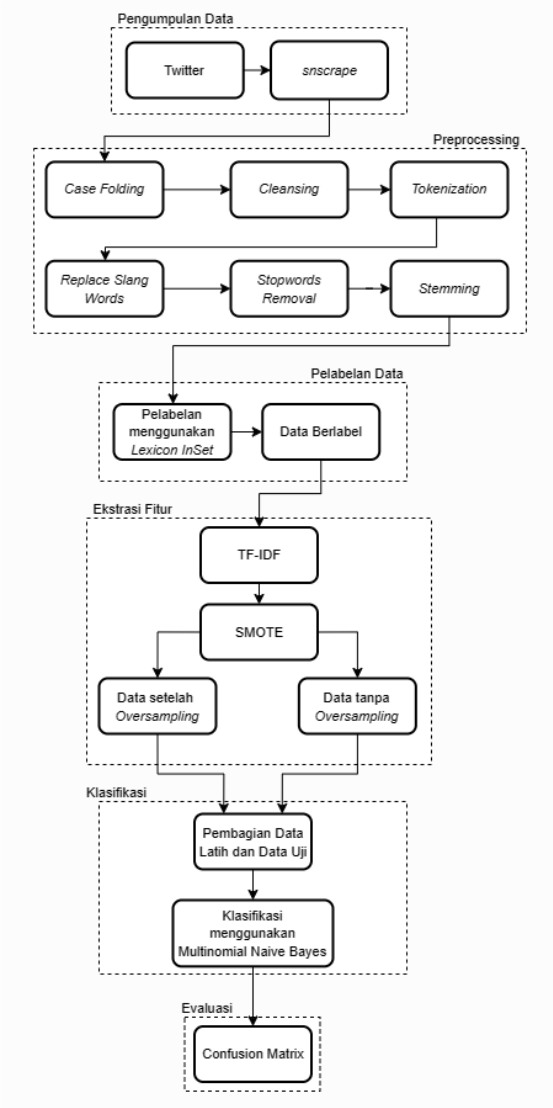
Data mining merupakan proses menemukan pola secara otomatis atau semiotomatis pada sekumpulan data yang besar yang disimpan pada *database*, data *warehouses*, atau penyimpanan informasi lainnya [23], [24]. Data mining dapat diklasifikasikan kedalam 2 kategori: deskriptif dan prediktif. Data mining deskriptif berguna untuk menggambarkan sifat umum dari data dalam database. Beberapa teknik yang digunakan dalam data mining deskriptif seperti *clustering* dan *association rules*. Data mining prediktif melakukan penarikan kesimpulan pada data saat ini untuk membuat prediksi[24].Beberapa teknik yang digunakan dalam data mining prediktif ini seperti: klasifikasi dan regresi.

2.8. *Confusion Matrix*

Confusion matrix menunjukkan jumlah prediksi yang benar dan salah yang dibuat oleh model saat memprediksi kelas target pada suatu dataset. Ini adalah alat yang berguna untuk melakukan pengukuran terhadap kinerja model klasifikasi. Terdapat 4 komponen pada *confusion matrix* yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN).

3. METODE PENELITIAN

Berikut merupakan tahapan metode penelitian yang dilakukan



Gambar 1. Metode Penelitian

3.1. Metode Pengumpulan Data

Pengumpulan data pada Twitter dilakukan menggunakan *library snscrape*. *Snscrape* merupakan *scrapper* untuk *social networking services* (SNS) seperti Facebook, Instagram, Twitter dan lain-lain. Data diambil dengan kata kunci "anies calon presiden", "prabowo calon presiden" pada saat keduanya dicalonkan menjadi bakal calon presiden oleh suatu partai. Data Anies di *crawling* dari tanggal 3 Oktober 2022 sampai 3 November 2022 dan Data Prabowo di *crawling* dari tanggal 12 Agustus 2022 sampai 12 September 2022.

3.2. *Preprocessing*

Setelah melakukan pengumpulan data, terdapat beberapa data yang tidak relevan seperti iklan, yang dimana ini akan dihapus terlebih dahulu sebelum data dilakukan *preprocessing*. Proses penghapusan data tersebut memakai cara manual pada Excel dengan menganalisa setiap datanya. Selanjutnya, data masuk ke tahap *preprocessing* data, yang dimana dari data tweets yang didapat sebelumnya akan dilakukan proses *case folding*, *cleansing*, *tokenization*, *replace slang word*, *stopword removal*, dan *stemming*.

3.3. Pelabelan Data

Setelah data telah melalui tahap *preprocessing* maka akan dilanjutkan dengan pelabelan data, yang dilakukan dengan metode *lexicon-based*. *Lexicon* yang digunakan adalah *Indonesia Sentiment Lexicon* (InSet). *Lexicon* InSet ini memiliki kamus berbahasa Indonesia yang berjumlah sekitar 10.000 kata. Setiap kata memiliki bobot yang berkisar dari angka -5 (sangat negatif) sampai dengan 5 (sangat positif)[18].

3.4. Ekstraksi Fitur

Pembobotan dengan *Term Frequency-Inverse Document Frequency* (TF-IDF) berguna untuk mengekstraksi kata inti dari dokumen, menghitung derajat yang sama diantara dokumen, menentukan peringkat pencarian[22]. Kata *term frequency* (TF) berarti kemunculan kata tertentu dalam suatu dokumen. Sedangkan untuk *inverse document frequency* (IDF) digunakan untuk mengukur seberapa penting suatu kata dalam seluruh dokumen. Pada penelitian ini akan menggunakan 2 skema ukuran n-gram diantaranya 1-gram dan 2-gram.

3.5. Pembangunan Model Klasifikasi

Setelah data melalui tahap pelabelan dan ekstrasi fitur, maka data akan dilanjutkan kedalam proses klasifikasi menggunakan algoritma *Multinomial Naïve Bayes*. *Multinomial NB* ini memiliki kesamaan dengan *Naïve Bayes* yang sama-sama mempertimbangkan metode probabilistik [8]. Terdapat dua tahap klasifikasi dokumen yang digunakan pada penelitian ini. Pertama, melakukan pelatihan terhadap dokumen yang kelasnya sudah diketahui. Kedua, merupakan proses klasifikasi dokumen yang belum diketahui kelasnya. Pengujian data dibagi menjadi dua skenario yakni dengan skenario 70% data latih dan 30% data uji, 80% data latih dan 20% data uji.

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

Data yang digunakan merupakan data pada media sosial Twitter yang telah di-crawling menggunakan library snscrape. Terdapat 2 bakal calon presiden yang digunakan yakni Anies dan Prabowo

yang di-crawling pada saat keduanya dicalonkan oleh partai yang mendukungnya. Data Anies diambil dengan kata kunci "anies calon presiden" pada tanggal 3 Oktober sampai 3 November 2022 yang berjumlah sebanyak 1.106 tweets. Data Prabowo diambil dengan kata kunci "prabowo calon presiden" pada tanggal 12 Agustus sampai 12 September 2022, yang berjumlah sebanyak 1.762 tweets. Data ini terdapat 3 atribut yakni *id*, *date*, dan *tweet*. Penjelasan mengenai atribut yang terdapat pada dataset ditunjukkan pada Tabel 1.

Tabel 1. Penjelasan Atribut Pada Dataset

No	Atribut	Deskripsi
1.	id	Berisi id dari pengguna Twitter.
2.	date	Memuat informasi tentang waktu pada saat pengguna mengirim *tweet*.
3.	tweet	Memuat opini yang diberikan oleh pengguna Twitter.

Adapun contoh dari data tweet yang berhasil di crawling ditunjukkan pada Tabel 2.

Tabel 2. Contoh Data Tweet

No	Tweet
1.	@ChusnulCh__ @gembong_warsono Tapi Anies Baswedan adalah Calon Presiden 'terkuat' untuk memimpin RI di 2024....
2.	@azwarsiregar Saya dukung sepenuhnya Pak Prabowo menjadi Calon Presiden. Semoga sukses.

4.2. *Preprocessing*

Setelah melakukan pengumpulan data, maka dilanjutkan ke tahap *preprocessing*. Tahapan ini akan melakukan perubahan terhadap data teks yang mentah yang didapat pada proses crawling menjadi data yang 'bersih' agar data teks dapat diproses lebih lanjut. Selain itu juga pada tahap ini akan dilakukan pemilihan atribut yang akan digunakan. Adapun tahapan *preprocessing* data yang dilakukan pada penelitian ini meliputi proses *case folding*, *cleansing*, *tokenization*, *replace slang word*, *stopword removal*, dan *stemming*. Tahapan *case folding* dilakukan untuk mengubah huruf kapital menjadi huruf kecil, proses ini ditunjukkan pada Tabel 3.

Tabel 3. Proses *Case Folding*

Tweet	*Case Folding*
@ChusnulCh__ @gembong_warsono Tapi Anies Baswedan adalah Calon Presiden 'terkuat' untuk memimpin RI di 2024....	@chusnulch__ @gembong_warsono tapi anies baswedan adalah calon presiden 'terkuat' untuk memimpin ri di 2024....
@azwarsiregar Saya dukung sepenuhnya Pak Prabowo menjadi Calon Presiden. Semoga sukses.	@azwarsiregar saya dukung sepenuhnya pak prabowo menjadi calon presiden. semoga sukses.

Selanjutnya tahapan *cleansing* dilakukan untuk menghapus karakter atau elemen yang tidak revelan pada data *tweet*, seperti simbol atau karakter numerik, tanda baca, *link URL*, emoji, *mention*, *retweet*, dan *hashtag*. Proses ini ditunjukkan pada Tabel 4.

Tabel 4. Proses *Cleansing*

Tweet	*Cleansing*
@chusnulch__ @gembong_warsono tapi anies baswedan adalah calon presiden 'terkuat' untuk memimpin ri di 2024....	tapi anies baswedan adalah calon presiden terkuat untuk memimpin ri di
@azwarsiregar saya dukung sepenuhnya pak prabowo menjadi calon presiden. semoga sukses.	saya dukung sepenuhnya pak prabowo menjadi calon presiden semoga sukses

Pada tahap selanjutnya dilakukan *tokenization*, untuk kalimat pada tweet menjadi potongan kata-kata yang disebut dengan *token*. Tahapan ini ditunjukkan pada Tabel 5.

Tabel 5 Proses *Tokenization*

Tweet	*Tokenization*
[tapi, anies, baswedan, adalah, calon, presiden, terkuat, untuk, memimpin, ri, di]	[tapi, anies, baswedan, adalah, calon, presiden, terkuat, untuk, memimpin, ri, di]
[saya, dukung, sepenuhnya, pak, prabowo, menjadi, calon, presiden, semoga, sukses]	[saya, dukung, sepenuhnya, pak, prabowo, menjadi, calon, presiden, semoga, sukses]

Pada tahap *replace slang words* akan mengubah kata-kata yang tidak formal atau kata gaul pada data tweet menjadi kata formal. Daftar kata tidak formal atau gaul diperoleh dari dataset yang ada dan dikombinasikan dengan kamus *Colloquial Indonesian Lexicon*.

Setelah itu pada tahap *stopwords removal* bertujuan untuk menghapus kata-kata umum dan sering digunakan, seperti: dan, atau, dan lain-lain. Kamus stopwords yang digunakan adalah kamus stopwords bahasa Indonesia. Proses *stopwords removal* ini ditunjukkan pada Tabel 6.

Tabel 6 Stopwords Removal

Tweet	*Stopwords Removal*
[tapi, anies, baswedan, adalah, calon, presiden, terkuat, untuk, memimpin, ri, di]	[anies, baswedan, calon, presiden, terkuat, memimpin, ri]
[saya, dukung, sepenuhnya, pak, prabowo, menjadi, calon, presiden, semoga, sukses]	[dukung, sepenuhnya, pak, prabowo, menjadi, calon, presiden, semoga, sukses]

Selanjutnya pada tahapan terakhir dilakukan *stemming*, yang melakukan perubahan terhadap kata imbuhan menjadi kata dasar yang sesuai dengan kaidah Kamus Besar Bahasa Indonesia. Proses ini menggunakan *library Sastrawi*. Proses *Stemming* ini ditunjukkan pada Tabel 7.

Tabel 7. Proses Stemming

Tweet	*Stemming*
[anies, baswedan, calon, presiden, terkuat, memimpin, ri]	[anies, baswedan, calon, presiden, kuat, pimpin, ri]
[dukung, sepenuhnya, pak, prabowo, menjadi, calon, presiden, semoga, sukses]	[dukung, sepenuh, pak, prabowo, jadi, calon, presiden, moga, sukses]

4.3. Pelabelan Data

Pada tahapan ini data akan dilakukan proses pelabelan dengan menggunakan *lexicon* InSet. Dalam suatu kalimat bisa mendapatkan nilai 0 sampai 5 yang menyatakan ada atau tidaknya sentimen yang terkandung. *Lexicon* InSet menggunakan dua nilai yang disebut InSetPos dan InSetNeg. InSetPos menyatakan jumlah nilai untuk kata-kata positif sedangkan InSetNeg menyatakan jumlah nilai untuk kata-kata negatif. Setelah itu kedua nilai ditambahkan untuk mendapatkan nilai akhir untuk mendapatkan polaritas sentimennya. Hasil dari pelabelan ditunjukkan pada Tabel 8.

Tabel 8. Hasil Pelabelan Data

Dataset	Label	
	Positif	Negatif
Anies	368	172
Prabowo	420	125

Dari hasil pelabelan yang ditunjukkan pada Tabel 8 terlihat adanya ketidakseimbangan terhadap kelas data yang ditunjukkan dengan adanya satu kelas yang lebih dominan dibanding kelas lainnya.

4.4. Ekstraksi Fitur

Pada tahap ini, data akan dilakukan pembobotan dengan TF-IDF. *Term Frequency* (TF) berarti kemunculan kata tertentu dalam suatu dokumen. *Inverse Document Frequency* (IDF) digunakan untuk mengukur seberapa unik dan penting suatu kata. Pada tahap ini juga akan dilakukan dua skenario penggunaan n-gram yakni, 1-gram dan 2-gram. Setelah itu data akan dilakukan teknik *oversampling* SMOTE.

4.5. Pembuatan Model Klasifikasi

Pada tahap ini, dilakukan pembuatan model dengan algoritma Multinomial NB, yang dimana akan melakukan perhitungan dengan model algoritma tersebut untuk melakukan klasifikasi terhadap data berdasarkan data latih yang telah memiliki kelas. Untuk pembagian data latih dan data uji menggunakan dua skenario yakni 70%:30% dan 80%:20%. Pada tahap ini juga akan dilakukan eksperimen untuk melakukan klasifikasi pada data dengan kelas yang tidak seimbang dan data dengan kelas yang seimbang. Hasil dari penelitian dibagi menjadi 2 bagian, yaitu hasil pengujian dengan kelas data tidak seimbang dan kelas data seimbang dan hasil pengujian dengan 1-gram dan 2-gram.

4.6. Hasil Pengujian Dengan Kelas Data Tidak Seimbang dan Kelas Data Seimbang

Hasil dari pengujian dengan kelas data tidak seimbang dan kelas data seimbang dilakukan dengan metode *confusion matrix*. Adapun hasil pengujian ditunjukkan pada Tabel 9.

Tabel 9. Pengujian pada kelas data seimbang dan tidak seimbang

Metode	Split Data	Akurasi	
		Dataset Anies	Dataset Prabowo
MNB	70%:30%	65,43%	83,53%
MNB+SMOTE	70%:30%	85,05%	90,47%
MNB	80%:20%	66,66%	81,65%
MNB+SMOTE	80%:20%	87,16%	89,28%

Dari Tabel 9 dapat diketahui untuk nilai akurasi untuk kedua data cenderung meningkat jika kelas data telah seimbang. Pada dataset Anies nilai akurasi tertinggi didapat pada pembagian data 80% data latih dan 20% data uji yang mendapatkan akurasi 87,16%. Pada dataset Prabowo nilai akurasi tertinggi didapat pada pembagian data 70% data latih dan 30% data uji yang mendapatkan nilai akurasi 90,47%.

4.7. Hasil Pengujian Dengan 1-gram dan 2-gram

Penggunaan 1-gram dan 2-gram dilakukan pada tahap TF-IDF, dan dilanjutkan dengan proses klasifikasi. Hasil dari pengujian ditunjukkan pada Tabel 10.

Tabel 10 Pengujian N-gram pada Dataset Anies

Metode	*Dataset*	Jumlah Fitur	*Accuracy*	*Precision*
1-gram	Tidak Seimbang	1.325	66,66%	44,44%
1-gram	Seimbang	1.325	87,16%	88,45%
2-gram	Tidak Seimbang	3.708	66,66%	44.44%
2-gram	Seimbang	3.708	87,16%	87,79%

Pada pengujian N-gram pada dataset Anies, menunjukkan hasil nilai akurasi yang baik didapat jika data seimbang dibanding dengan data yang tidak seimbang. Pada pengujian ini memiliki nilai akurasi yang cenderung sama, namun adanya perbedaan pada nilai *precision*. Nilai precision tertinggi didapatkan dengan masukan 1-gram ketika kelas data telah seimbang dengan menggunakan oversampling SMOTE, yakni sebesar 88,45%.

Tabel 11. Pengujian N-gram pada Dataset Prabowo

Metode	*Dataset*	Jumlah Fitur	*Accuracy*	*Precision*
1-gram	Tidak Seimbang	1.325	81,65%	66,66%
1-gram	Seimbang	1.325	89,28%	90,76%
2-gram	Tidak Seimbang	3.708	81,65%	66,66%
2-gram	Seimbang	3.708	94,04%	94,45%

Pada Tabel 11 menunjukkan pengujian N-gram dengan pada dataset Prabowo. Hasilnya menunjukkan bahwa hasil akurasi berkisar diantara 94,04% hingga 81,65%. Nilai akurasi tertinggi didapatkan dengan masukan 2-gram dengan nilai akurasi sebesar 94,04% dengan kelas data yang seimbang. Dari hasil tersebut menunjukkan penggunaan 2-gram pada dataset ini menghasilkan performa akurasi yang lebih baik dalam melakukan klasifikasi dibanding dengan penggunaan 1-gram, karena penggunaan 2-gram dapat memberikan informasi yang lebih baik karena mempertimbangkan pasangan kata yang berdekatan dengan fitur.

Untuk hasil pengujian yang digunakan pada penelitian ini adalah hasil pengujian yang mendapatkan nilai *accuracy* yang baik pada tiap pengujian yang telah dilakukan.

5. KESIMPULAN DAN SARAN

Berdasarkan pada penelitian yang telah dilakukan pada data sentimen masyarakat di Twitter terhadap bakal calon presiden 2024, algoritma Multinomial Naïve Bayes dengan metode oversampling SMOTE dapat diterapkan untuk analisis sentimen dan menunjukkan hasil yang cukup baik. Penggunaan metode oversampling SMOTE terbukti dapat meningkatan performa klasifikasi jika terjadi ketidakseimbangan kelas data, yang dimana dibuktikan dengan adanya peningkatan akurasi pada setiap dataset yang digunakan. Pada dataset Anies dengan kelas data yang tidak seimbang mendapatkan akurasi berkisar dari 65,43% sampai 66,66%. Setelah dataset Anies di sintesis menggunakan metode oversampling SMOTE hasil akurasi sebelumnya meningkat berkisar antara 90,47% dan 89,28%. Dari hasil tersebut terlihat adanya peningkatan sebesar sekitar 20% untuk dataset Anies ketika kelas data telah seimbang. Sedangkan pada dataset Prabowo dengan kelas data yang tidak seimbang mendapatkan akurasi berkisar dari 83,53% sampai 81,65%. Ketika dataset tersebut di sintesis menggunakan metode oversampling SMOTE hasilnya meningkat pada kisaran 90,47% dan 89,28%. Dari hal tersebut, terlihat adanya peningkatan sekitar 6-7% untuk dataset Prabowo ketika kelas data telah seimbang. Penggunaan N-gram pada kedua dataset menghasilkan hasil yang berbeda-beda karena setiap dataset memiliki karakteristiknya masing-masing. Pada dataset Anies didapatkan hasil *accuracy* dan *precision* yang baik dengan menggunakan 1-gram, sedangkan dataset Prabowo didapatkan nilai *accuracy* dan *precision* yang baik dengan menggunakan 2-gram. Selain itu, dapat dilakukan penelitian lebih lanjut untuk dengan menggunakan ukuran n-gram lebih dari 2 dan menggunakan jumlah data yang lebih banyak lagi.

DAFTAR PUSTAKA

[1] A. D. Akmal, I. Permana, H. Fajri, and Y. Yuliarti, "Opini Masyarakat Twitter terhadap Kandidat Bakal Calon Presiden Republik Indonesia Tahun 2024," *Jurnal Manajemen dan Ilmu Administrasi Publik (JMIAP)*, vol. 4, no. 4, pp. 292–300, Dec. 2022, doi: 10.24036/jmiap.v4i4.160.

[2] Galih Wicaksono, "Euforia Pemilu 2024, Kembali Menggelora dan Lahirnya Poros Baru Koalisi Kebangkitan Indonesia Raya," Aug. 09, 2022. https://www.kompasiana.com/galihwicaksono6461/62f206a2a51c6f2f6d370f32/euforia-pemilu-2024-kembali-menggelora-dan-lahirnya-poros-baru-koalisi-kebangkitan-indonesia-raya?page=1&page_images=1 (accessed Jan. 15, 2023).

[3] Edwin Zhan, "Charta Politika Rilis Survei Elektabilitas Terbaru untuk Capres 2024: Ganjar-Prabowo-Anies Tertinggi," May 15, 2023. https://www.kompas.tv/article/407042/charta-politika-rilis-survei-elektabilitas-terbaru-untuk-capres-2024-ganjar-prabowo-anies-tertinggi (accessed May 21, 2023).

[4] Dian Erika Nugraheny, "Survei SMRC: Elektabilitas Ganjar 24,6 Persen, Prabowo 17,2 Persen, Anies 10,7 Persen," May 12, 2023. https://nasional.kompas.com/read/2023/05/12/19241181/survei-smrc-elektabilitas-ganjar-246-persen-prabowo-172-persen-anies-107 (accessed May 21, 2023).

[5] Simon Kemp, "DIGITAL 2023: INDONESIA," Feb. 09, 2023. https://datareportal.com/reports/digital-2023-indonesia (accessed Jun. 08, 2023).

[6] P. O. Abas Sunarya, R. Refianti, A. B. Mutiara, and W. Octaviani, "Comparison of Accuracy between Convolutional Neural Networks and Naïve Bayes Classifiers in Sentiment Analysis on Twitter," 2019. [Online]. Available: www.ijacsa.thesai.org

[7] H. M. Ismail, S. Harous, and B. Belkhouche, "A Comparative Analysis of Machine Learning Classifiers for Twitter Sentiment Analysis," *Research in Computing Science*, vol. 110, no. 1, pp. 71–83, Dec. 2016, doi: 10.13053/rcs-110-1-6.

[8] M. B. Ressan and R. F. Hassan, "Naïve-Bayes family for sentiment analysis during COVID-19 pandemic and classification tweets," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 1, pp. 375–383, Oct. 2022, doi: 10.11591/ijeecs.v28.i1.pp375-383.

[9] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Intern J Pattern Recognit Artif Intell*, vol. 23, no. 4, pp. 687–719, Jun. 2009, doi: 10.1142/S0218001409007326.

[10] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," 2002.

[11] W. Satriaji and R. Kusumaningrum, "Effect of Synthetic Minority Oversampling Technique

(SMOTE), Feature Representation, and Classification Algorithm on Imbalanced Sentiment Analysis," in *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS)*, IEEE, Oct. 2018, pp. 1–5. doi: 10.1109/ICICOS.2018.8621648.

[12] S. Kumar, A. K. Kar, and P. V. Ilavarasan, "Applications of text mining in services management: A systematic literature review," *International Journal of Information Management Data Insights*, vol. 1, no. 1. Elsevier Ltd, Apr. 01, 2021. doi: 10.1016/j.jjimei.2021.100008.

[13] M. D. Devika, C. Sunitha, and A. Ganesh, "Sentiment Analysis: A Comparative Study on Different Approaches," in *Procedia Computer Science*, Elsevier B.V., 2016, pp. 44–49. doi: 10.1016/j.procs.2016.05.124.

[14] V. Bonta, N. Kumaresh, and N. Janardhan, "A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis," *Asian Journal of Computer Science and Technology*, vol. 8, no. S2, pp. 1–6, Mar. 2019, doi: 10.51983/ajcst-2019.8.s2.2037.

[15] N. L. Octaviani, E. Hari Rachmawanto, C. A. Sari, and D. Rosal Ignatius Moses Setiadi, "Comparison of multinomial naïve bayes classifier, support vector machine, and recurrent neural network to classify email spams," in *Proceedings - 2020 International Seminar on Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, iSemantic 2020*, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 17–21. doi: 10.1109/iSemantic50169.2020.9234296.

[16] C. D. Manning, Prabhakar. Raghavan, and Hinrich. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2008.

[17] Z. Drus and H. Khalid, "Sentiment analysis in social media and its application: Systematic literature review," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 707–714. doi: 10.1016/j.procs.2019.11.174.

[18] F. Koto and G. Y. Rahmaningtyas, "Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs," in *2017 International Conference on Asian Language Processing (IALP)*, IEEE, Dec. 2017, pp. 391–394. doi: 10.1109/IALP.2017.8300625.

[19] J. Ah-Pine, E. Pavel, and S. Morales, "A Study of Synthetic Oversampling for Twitter Imbalanced Sentiment Analysis." [Online]. Available: http://www.internetlivestats.com/twitter-statistics/

[20] J. A. Septian, M. T. Fahrudin, and A. Nugroho, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor," *JOURNAL OF INTELLIGENT SYSTEMS AND COMPUTATION*, vol. 1, pp. 43–49, Aug. 2019, doi: https://doi.org/10.52985/insyst.v1i1.36.

[21] G. Paltoglou and M. Thelwall, "A study of Information Retrieval weighting schemes for sentiment analysis," Association for Computational Linguistics, 2010.

[22] S. W. Kim and J. M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, Dec. 2019, doi: 10.1186/s13673-019-0192-7.

[23] H. I. Witten, E. Frank, and A. H. Mark, *Data Mining Practical Machine Learning Tools and Techniques*, vol. Third Edition. 2011.

[24] H. Jiawei and Micheline Kamber, *Data Mining: Concepts and Techniques*.